D-ESRGAN: A Dual-Encoder GAN with Residual CNN and Vision Transformer for Iris Image Super-Resolution

Caiyong Wang[1,*], Tianhao Lu[2,†], Gaosheng Wu[3], Yunlong Wang[2], Zhenan Sun[2]

[1]School of Electrical and Information Engineering,
Beijing University of Civil Engineering and Architecture, Beijing, P.R. China
[2]CRIPAC, NLPR, CASIA, Beijing, P.R. China [3]Baidu Inc., Beijing, P.R. China

[*]wangcaiyong@bucea.edu.cn, [†]Tianhao.lu@cripac.ia.ac.cn(corresponding author)

Abstract

Iris images captured in less-constrained environments, especially at long distances often suffer from the interference of low resolution, resulting in the loss of much valid iris texture information for iris recognition. In this paper, we propose a dual-encoder super-resolution generative adversarial network (D-ESRGAN) for compensating texture lost of the raw image meanwhile maintaining the newly generated textures more natural. Specifically, the proposed D-ESRGAN not only integrates the residual CNN encoder to extract local features, but also employs an emerging vision transformer encoder to capture global associative information. The local and global features from two encoders are further fused for the subsequent reconstruction of high-resolution features. During the training, we develop a three-stage strategy to alleviate the problem that generative adversarial networks are prone to collapse. Moreover, to boost the iris recognition performance, we introduce a triplet loss to push away the distance of super-resolved iris images with different IDs, and pull the distance of super-resolved iris images with the same ID much closer. Experimental results on the public CASIA-Iris-distance and CASIA-Iris-M1 datasets show that D-ESRGAN archives better performance than state-of-the-art baselines in terms of both super-resolution image quality metrics and iris recognition metric.

1. Introduction

Since the iris contains distinctive texture details, which are suitable for biometric identification, iris recognition is widely applied in various scenarios, such as access control, mobile payment, forensics. However, when iris images are captured in less-constrained environments, especially at long distances, the image quality is significantly degenerated. Specially, the image resolution becomes small and the clear texture details become blurred, which seriously degrades the recognition performance of the existing iris recognition system. In this case, we either have to upgrade the existing iris recognition system to adapt to the low-quality and low-resolution iris image, or consider the enhancement of the low-quality and low-resolution iris image to meet the quality requirements of the existing system. Compared with the former, the latter is less costly and has therefore drawn considerable research interests recently. Among the many iris image enhancement technologies, iris super-resolution is an effective method that can map low-quality and low-resolution iris images to high-quality and high-resolution iris images, as shown in Figure 1.

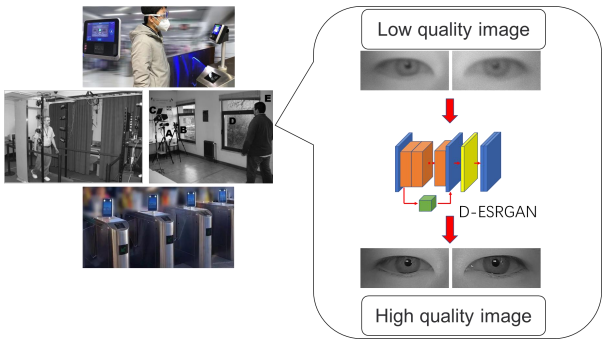


Figure 1. Less-constrained iris recognition scenarios where the proposed iris super-resolution method is used to enhance the iris image.

Traditional image super-resolution methods, *e.g.*, nearest neighbor interpolation [17], triangular mesh method [12] are prone to smooth out and blur the image. By contrast, the deep learning based approaches have shown the impressive results. SRCNN [2] was the first image super-resolution method based on convolutional neural networks (CNNs), which included three modules, *i.e.*, patch extraction and representation, non-linear mapping, and feature reconstruc-

tion, and eventually achieved better super resolution results than traditional methods. Later improvement methods include VDSR [9], DnCNN [28], and IRCNN [29]. These methods mainly structurally enhance the image after super resolution, however the texture is still relatively blurry. To solve this problem, SRGAN [14] employed a generative adversarial network (GAN) framework, where multiple residual CNN modules were first stacked in the generator to improve the detail perception ability of the model, then the discriminator was employed to make the generated image textures more natural by an adversarial game with the generator. Furthermore, to alleviate the problem of jagged edges in the generated images, ESRGAN [23] removed the batch normalization layer to increase the richness of middle features, and performed the supervised learning at the feature layer instead of the image pixel layer.

Recently, Transformer [21] has made great success in various NLP tasks due to its superior ability to capture long range dependencies and model the contextual information via attention mechanism. Encouraged by these benefits, vision transformer has also been proposed to improve the performance of the computer vision (CV) tasks, including image super-resolution. TTSR [25] is a novel texture transformer network for image super-resolution, where the low-resolution image and high-resolution reference image are formulated as queries and keys in a transformer, respectively. By jointly learning the features across low-resolution and reference images, the deep feature correspondences can be discovered by attention, and thus accurate texture features can be transferred. Although this method is effective, it may be unsuitable for iris image super-resolution due to the lack of high-quality reference images with similar textures.

In this paper, we consider the benefits of convolutional neural network, vision transformer, and GAN simultaneously, and propose a dual-encoder super-resolution GAN (D-ESRGAN) model for iris image super-resolution. Specifically, the proposed D-ESRGAN model consists of a dual-encoder generator and a VGG-style discriminator. The generator not only integrates the residual CNN encoder to extract local features, but also employs an emerging vision transformer encoder to capture global positional associative information. Then these low-resolution features are fused and upsampled to a high-resolution size, and further reconstructed via several CNN blocks to generate the super-resolved iris images. The discriminator is responsible for dynamically supervising the training of the generator by learning the differences between the generated image and the ground-truth image at the feature level. Besides, a triplet loss and a three-stage strategy are proposed to assist in the training of the proposed model. Our main contributions can be summarized as follows:

- We propose a dual-encoder super-resolution GAN (D-ESRGAN) model for iris image super-resolution, where the generator in the GAN specially adopts a hybrid dual-encoder structure that employs the residual CNN layers to extract local features and the vision transformer layers to capture global positional associative information. This design retains the fast convergence property of CNN and has a global receptive field inherited from Transformer.

- During the model training, we introduce a triplet loss to push away the distance of super-resolved iris images with different IDs, and pull the distance of super-resolved iris images with the same ID much closer. This manner makes the generated super-resolution iris image textures more conducive to iris recognition.

- We develop a three-stage training strategy to stabilize the model training, where they focus on the feature reconstruction, image perception and iris recognition, respectively.

- Experimental results on the public CASIA-Iris-distance and CASIA-Iris-M1 datasets show that D-ESRGAN archives better performance than state-of-the-art baselines in terms of both super-resolution image quality metrics and iris recognition metric.

2. Related Work

In this section, we review previous image super-resolution methods which are the most relevant to our work.

Traditional methods. Yang *et al*. [27] adopted the sparse representation for image super-resolution, where images were represented as dictionaries and atoms, and dictionaries between high-quality and low-quality images were built. Gao *et al*. [6] proposed a local embedding based image super-resolution method with extensive manual parameter setting being required. Other traditional methods also include nearest neighbor interpolation [17], triangular mesh method [12], *etc*. In general, traditional methods tend to smooth out the image and do not portray edge information well.

CNN-based methods. SRCNN [2] was the first to introduce the CNN architecture into image super-resolution tasks and consisted of three modules, *i.e*., patch extraction and representation, non-linear mapping, and feature reconstruction. VDSR [9] introduced residual CNN blocks to learn the difference between paired high-resolution and low-resolution images efficiently and gained a better convergence speed during the training. Lim *et al*. [15] improved the ResNet-based super-resolution model by removing the Batch Normalization (BN) and ReLU layers, which constrained the diverse representability of the model and was detrimental to the super-resolution task. Furthermore,

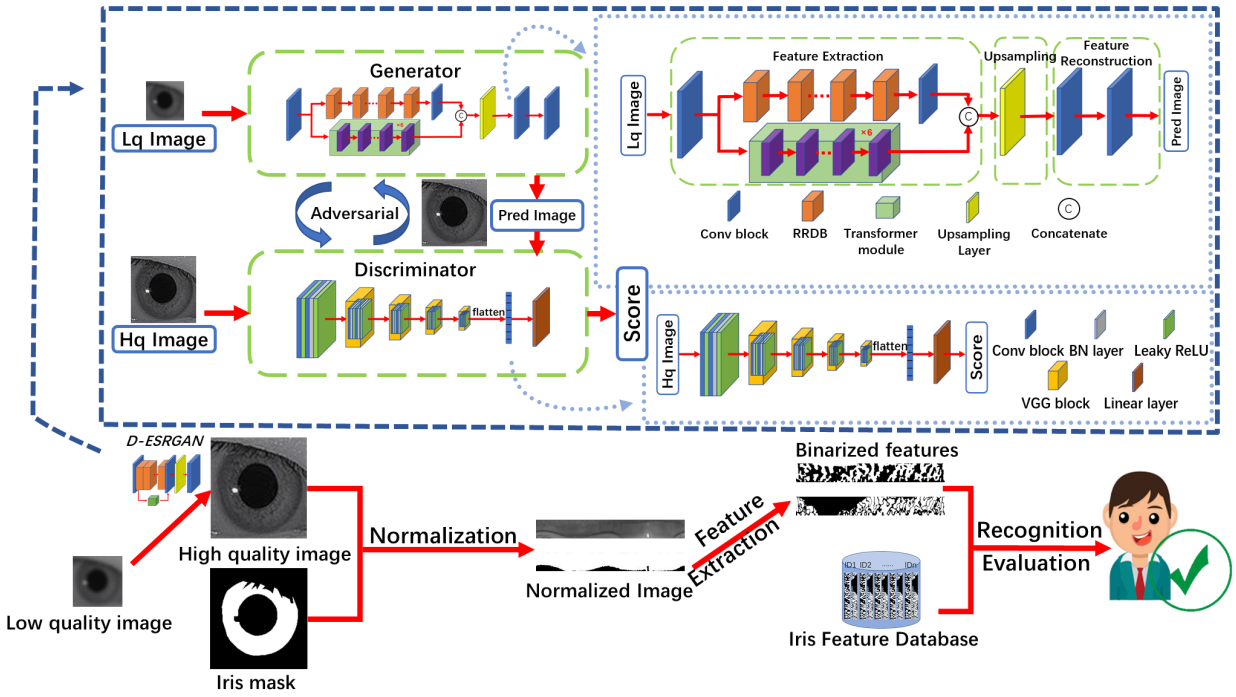


Figure 2. The detailed structure diagram of our proposed D-ESRGAN model, where the bottom shows the iris recognition pipeline with the super-resolved iris images generated by our method as the input.

attention mechanism [10, 30] and progressive reconstruction [5, 13] are introduced into the super-resolution network to enhance the image generation quality, respectively.

GAN-based methods. SRGAN [14] was the first to introduce the GAN into image super-resolution tasks and proposed the perceptual measure of feature-level distance and the pixel-level similarity loss to assist in the model training. ESRGAN [23] further improved SRGAN by removing the BN layer and adopting the residuals in the residual dense block (RRDB) to achieve a better receptive field. EnhanceNet [18] proposed a texture matching loss and built a pixel-level supervision to relieve the image smoothing problem.

Vision transformer-based methods. TTSR [26] was one of the first to introduce the transformer architecture into image super-resolution tasks. It adopted the high-resolution images as reference images to help the reconstruction of low-resolution images. By stacking multiple texture transformers with four closely-related modules, the model learned a more powerful feature representation and achieved a better super-resolution performance.

3. Approach

In this section, we introduce the proposed dual-encoder super-resolution GAN (D-ESRGAN) model for iris image super-resolution. As shown in Figure 2, D-ESRGAN comprises a dual-encoder generator based on residual CNN and vision transformer, and a VGG-style discriminator, which are described in Section 3.1 and Section 3.2, respectively. Then a group of loss functions including the triplet loss for optimizing the proposed model are presented in Section 3.3. Finally, we introduce an effective three-stage training strategy in Section 3.4.

More specifically, considering that a GAN consists of a generator and a discriminator, hence D-ESRGAN is formulated as follows:

$$\begin{aligned} \mathrm{x_h} &= G(x_l) \\ y_{dis} &= D(x_i), x_i \in \{x_h, x_{gt}\} \end{aligned} \tag{1}$$

where the generator G takes the low-resolution iris image x_l as input and generate a high-resolution iris image x_h. Then the discriminator D distinguishes the generated x_h to be fake ($y_{dis} = 0$) and the ground-truth high-resolution iris image x_{gt} to be real ($y_{dis} = 1$). When the adversarial training is performed, the generator tries to generate a more realistic high-resolution iris image, while the discriminator tries to determine whether the input image is real or not more confidently. Once the training iterations are enough, it is hard for the discriminator to distinguish the input to be real from fake, meanwhile the generator is able to generate a realistic enough high-resolution iris image. The trained generator is deployed for iris image super-resolution.

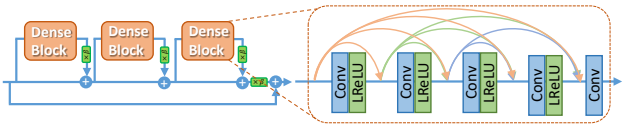


Figure 3. Residual-in-Residual Dense Block structure [23], where β is the residual scaling parameter.

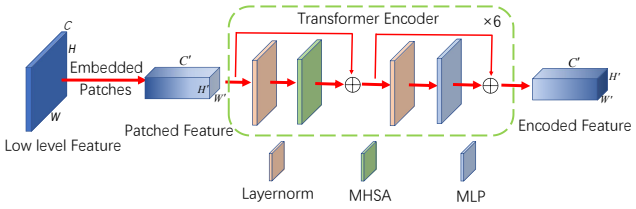


Figure 4. Vision transformer encoder structure.

3.1. Dual-Encoder Generator

The generator maps the low-resolution iris image into the high-resolution iris image by three modules, *i.e.*, low-resolution feature extraction, upsampling, and high-resolution feature reconstruction. The low-resolution feature extraction begins with a single convolutional block to extract the initial low-level features, then followed by the parallel residual CNN encoder and vision transformer encoder to extract the local and global features, respectively. In the following sections, we will describe the two encoders in detail.

3.1.1 Residual CNN Encoder

Following ESRGAN [23], we employ the Residual-in-Residual Dense Block (RRDB) without BN layers as the basic CNN encoder building unit. As shown in Figure 3, the RRDB combines multi-level residual network and dense block, which jointly improves the feature extraction ability of the encoder by enjoying the benefits of more layers and connections. Furthermore, BN layers are removed to reduce the computational complexity and help to improve generalization ability. In the residual network, a scaling parameter $\beta \in (0,\ 1)$ is applied before adding the dense block to prevent instability. Since the deeper network structure is beneficial to the super-resolution performance, the RRDB is stacked multiple times in the CNN encoder. Specifically, there are 23 RRDB blocks, and β is set to 0.2. Finally, a single convolutional block is further used to refine the local features from the residual CNN encoder. The whole process of the residual CNN encoder is formulated as follows:

$$x_{local} = f_1(\tilde{x_l}) \tag{2}$$

where $\tilde{x_l}$ denotes the initial low-level features of the input low-resolution iris image x_l, f_1 is the residual CNN encoder function that maps the initial low-level features to the refined local features x_{local}.

3.1.2 Vision Transformer Encoder

Although many texture information is lost in low-resolution images, there are still structural features in the image that can be utilized. However, the structural information in previously extracted low-level features are not adjacent in position and cannot be fully aggregated by modeling local features. Inspired by the vision transformer [4], we therefore introduce a vision transformer encoder to provide a global receptive field for capturing the positional associative information, as shown in Figure 4. It is worth noting that the encoder takes shallow low-level features as input, so that the original pixel blocks with similar basic textures are utilized for fusion prediction, which is more conducive to generating iris textures that can reflect the identity properties of individuals.

More specifically, we split the original sized feature map into multiple 8×8-sized feature patches, ensuring that all or some of the basic texture information is available in each feature patch. In the vision transformer encoder, we stack 6 alternating layers of Multi-Head Self-Attention (MHSA) and MLP blocks to model the long-range feature interactions and aggregate structurally related features, where Layernorm is applied before each block and residual connections are applied after each block. The whole process of the vision transformer encoder is formulated as follows:

$$x_{global} = f_2(\tilde{x_l}) \tag{3}$$

where f_2 is the vision transformer encoder function that maps the initial low-level features $\tilde{x_l}$ to the global features x_{global}.

After two encoders, we further concatenate their extracted features at the channel axis as the fused low-resolution features:

$$x_{fuse} = [x_{local}, x_{global}] \tag{4}$$

where x_{fuse} is the fused low-resolution features.

3.1.3 Upsampling and Feature Reconstruction

Inspired by the progressive up-sampling networks [13, 24], we progressively up-sample the low-resolution features x_{fuse} to a scaling factor. Taking 4 as an example, we zoom in the features by a factor of 2 each time, and thus the upsampling is performed twice. Here we adopt an efficient sub-pixel convolution (*i.e.*, pixel shuffle) [19] layer to learn to upscale the low-resolution features into the high-resolution features. Compared with the the handcrafted filters such as bilinear interpolation, bicubic interpolation, this manner allows for the specifical training of each

feature map, which makes the upsampled features more refined. Specifically, given the low-resolution feature $x_{fuse} \in R^{H\times W\times C}$ and the scaling factor of r, a 1×1 convolution layer first expands the channel dimension of x_{fuse} by a factor of r^2, yielding a new $x_{fuse} \in R^{H\times W\times C\cdot r^2}$, then followed by an periodic shuffling operator that rearranges the elements of the new x_{fuse} tensor to a upsampled feature tensor of shape $rH \times rW \times C$.

The upsampled features may contain some artifacts, hence it is necessary to perform a further feature reconstruction before the final prediction output. For this, two 1×1 convolution layers with ReLU activation are introduced to re-express the upsampled features and generate the final refined high-resolution image. The whole process is formulated as follows:

$$x_h = f_3(x_{fuse}) \tag{5}$$

where f_3 denotes the upsampling and feature reconstruction functions that map the fused low-resolution features x_{fuse} to the high-resolution image x_h.

3.2. VGG-style Discriminator

Following ESRGAN [23], we employ a VGG-style convolutional neural network as the discriminator. The VGG-style network contains five convolutional blocks with BN layer and Leaky ReLU activation layer, each reducing the feature map size to half. The obtained feature maps are further flatten, then followed by a fully connected layer to output the discriminant score, which indicates whether the input image is real or not.

3.3. Loss Function

The loss function of D-ESRGAN consists of 4 parts as following:

$$\begin{aligned} &L_{Total} = \\ &\lambda_1 L_{cont}(x_h, x_{gt}) + \lambda_2 L_{ad}(y_h, y_{gt}) \\ &+ \lambda_3 L_{per}(x_h, x_{gt}) + \lambda_4 L_{trip}(x_h, x_h^+, x_h^-), \end{aligned} \tag{6}$$

where x_h represents a high-resolution image generated by the generator. x_h^+ denotes other high-resolution sample image with the same identity ID as x_h, and x_h^- denotes other high-resolution sample image with a different identity ID from x_h. y_h and y_{gt} denote the score results obtained after feeding x_h and x_{gt} into the discriminator, respectively. $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ represent the weight coefficients, which are used to balance different loss functions in the training. L_{cont} denotes the structural loss, enabling the model to better learn iris contours, which is implemented as L_1 loss. L_{ad} denotes the generative adversarial loss [7], enabling the generator and discriminator to alternately optimize their respective model parameters in mutual supervision. L_{per} denotes the perceptual loss [23] based on pre-trained VGG19 features, making the generative results of the model more realistic and natural.

In order to better generate the super-resolved iris texture that characterizes identity information, we introduce an effective triplet loss as additional supervision:

$$L_{trip}(x_h, x_h^+, x_h^-) = \max(d(x_h, x_h^+) - d(x_h, x_h^-) + \alpha, 0) \tag{7}$$

where d denotes the Euclidean distance between two images in 2-dimensional space, and α is a parameter to control the desired margin between x_h-x_h^+ distance between x_h-x_h^- distance, which is set to 1.0 here. By optimizing the L_{trip}, the distance between samples with the same identity IDs will be reduced while the distance between samples with different identity IDs will be enlarged until their margin is larger than a certain value. Clearly, the L_{trip} forces the model to learn the common characteristics between the same identity ID images and the differences between different identity ID samples, hence the iris recognition information is encoded in the super-resolution model.

3.4. Three-stage Training Strategy

Stage	λ_1	λ_2	λ_3	λ_4	Learning Rate
stage 1	1	2e-2	0	0	2e-4
stage 2	1e-1	2e-2	1	0	2e-3
stage 3	1e-1	2e-2	2e-3	1	2e-3

Table 1. A summary of weight coefficients at each stage.

The whole training process is made of pre-training and fine-tuning. During the pre-training phase, we used the ND-IRIS-0405 [1] dataset composed of 64,980 iris samples from 356 subjects to train the initialization parameters of the model. During the fine-tuning phase, we adopt a three-stage strategy to train specific datasets. The different stages allow the model to focus on the optimization of different targets. In the first stage, L_{cont} and L_{ad} are employed to learn the rough iris contours for the initial feature reconstruction. In the second stage, L_{per} is newly added to make the super-resolution results more natural and realistic from the perspective of image perception. In the third stage, L_{trip} is further added to enable the model to learn the inter-class differences and intra-class commonalities of different ID iris images, so that the iris texture generated by the model maintains the identity information. The weight coefficients of the loss function at each stage are summarized in Table 1.

4. Experiments and Results

4.1. Experimental Setup

Datasets We conduct experiments using two publicly available iris datasets: CASIA-Iris-distance [20] and CASIA-Iris-M1 [20]. The first dataset was collected from

3 meters away with a CASIA long-range iris camera under near infrared illumination (NIR). For the super-resolution experiment, we use a subset from [16], which consists of 400 iris images with resolution of 640×480 pixels. The first 300 images are used for training, and the remaining 100 images are used for testing. The second dataset contains totally 11,000 images from 630 Asian subjects, which were collected using the mobile devices under NIR illumination. Similarly, we also use a subset from [22], which consists of 3,000 iris images with resolution of 400×400 pixels. The 1,500 iris images are randomly selected for training and the rest 1,500 iris images are used for testing. Both datasets are provided with the ground-truth iris masks. To unify two datasets in a framework, we manually extract the iris texture and iris mask with resolution of 224×224 pixels for each original image, which serves as the ground-truth high-resolution iris image. The low-resolution iris images are obtained by down-sampling the high-resolution iris images by 4 times using the bicubic interpolation, followed by a Gaussian blur to make the images more realistic. Therefore, all experiments are performed with a scaling factor of $\times 4$ between the low-resolution and high-resolution iris images.

Implementation details The proposed model is implemented in Pytorch on 4 NVIDIA 2080Ti GPUs. We use Adam [11] to optimise the model with a mini-batch size of 16 and L_2 weights regularization coefficient of 1×10^{-2} for 50 epoches, including 3 epoches for the first stage, 4 epoches for the second stage, and 43 epoches for the third stage. The learning rate of the generator at each stage are summarized in Table 1, and the learning rates of the discriminator are set to 0.01 times that of the generator. To evaluate the iris recognition performance, we conduct the recognition experiment on the testing set. Specifically, we firstly perform the iris super-resolution for the low-resolution iris images using the trained model. Then the generated iris images and their ground-truth iris masks are normalized, followed by the iris feature extraction using Uninet [31]. Finally, we perform the iris matching using the binarized iris features, as illustrated in Figure 2. For comparison, we select several baseline methods to reproduce, including traditional methods (*i.e.*, Linear interpolation, Nearest interpolation, and Bicubic interpolation), and deep learning-based methods (*i.e.*, SRCNN [3], SRGAN [14], ESRGAN [23], and IrisDNet [8]).

Evaluation protocols For the iris super-resolution task, we employ the Peak Signal to Noise Ratio (PSNR) and SSIM to evaluate the quality of the generated iris images. For the iris recognition task, we employ the equal error rate (EER) to evaluate the recognition performance of the super-resolved iris images.

Datasets	Methods	PSNR	SSIM	EER
CASIA-Iris-distance	Linear	33.5241	0.8424	0.3705
	Nearest	35.2714	0.8474	0.3419
	Bicubic	35.2953	0.8676	0.3581
	SRCNN [3]	35.2733	0.8487	0.3560
	SRGAN [14]	36.7358	0.9027	0.3360
	ESRGAN [23]	36.4151	0.8708	0.2683
	IrisDNet [8]	35.5642	0.9107	0.3410
	D-ESRGAN (ours)	**38.6067**	**0.9335**	**0.2333**
CASIA-Iris-M1	Linear	34.2797	0.8762	0.4414
	Nearest	36.4083	0.8818	0.4405
	Bicubic	36.4083	0.8911	0.4408
	SRCNN [3]	36.4083	0.8935	0.4405
	SRGAN [14]	38.3985	0.9320	0.4113
	ESRGAN [23]	40.1644	0.9542	0.3975
	IrisDNet [8]	40.0697	0.9616	0.4102
	D-ESRGAN (ours)	**41.5730**	**0.9559**	**0.3509**

Table 2. A quantitative comparison of different super-resolution methods on the image quality and iris recognition performance.

Methods	PSNR	SSIM	EER
Residual CNN Encoder	36.4151	0.8708	0.2683
+Vision Transformer Encoder	38.4267	0.9146	0.2530
+Vision Transformer Encoder+Triplet loss	38.5725	0.9287	0.2467
+Vision Transformer Encoder+Triplet loss+Pre-training	38.6067	0.9335	0.2333

Table 3. Results of different ablation settings.

4.2. Experimental Analysis

Quality Analysis The quantitative comparison between the proposed D-ESRGAN and baselines on the super-resolution image quality is shown in Table 2. Furthermore, the qualitative comparison is shown in Figure 5. From the quantitative and qualitative results, we can see that GAN-based methods (SRGAN, ESRGAN, and IrisDNet) exhibit clear advantages compared to traditional methods and early CNN-based SRCNN method in terms of PSNR and SSIM metrics. Moreover, our proposed D-ESRGAN consistently achieves the best results on the two datasets. It can better characterize iris texture details than other baseline methods.

Recognition Analysis The iris recognition results are shown in Table 2, and we also draw the corresponding DET curves in Figure 6. It can be seen that GAN-based methods (SRGAN, ESRGAN, and IrisDNet) show significant advantages than traditional methods and early CNN-based SRCNN method in terms of EER and DET curve. Our proposed D-ESRGAN further achieves the best recognition performance on the two datasets, respectively, which reflects that our super-resolution method is beneficial to improving the iris recognition performance.

Ablation Study To investigate the effectiveness of the proposed model components, we designed ablation experiments on the CASIA-Iris-distance dataset, and the results are shown in Table 3. From the results, we can see that: 1) the largest performance improvements are from the Vision Transformer Encoder, which shows that incorporating a global receptive field to capture the positional associative information is essential for promoting the iris super-

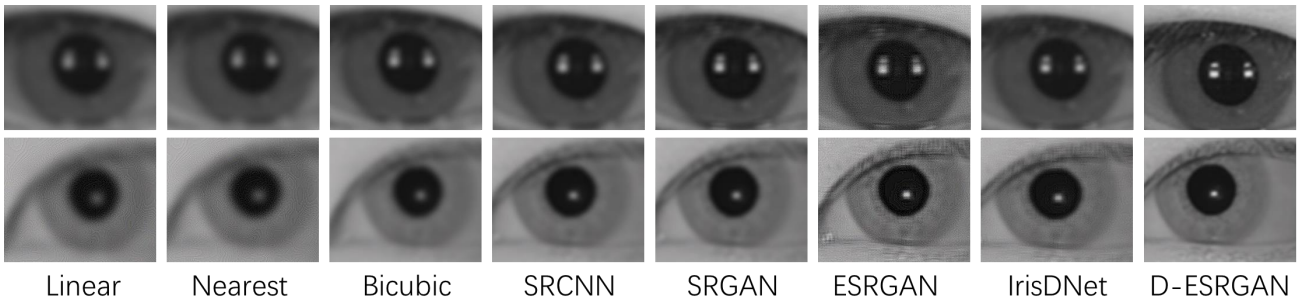
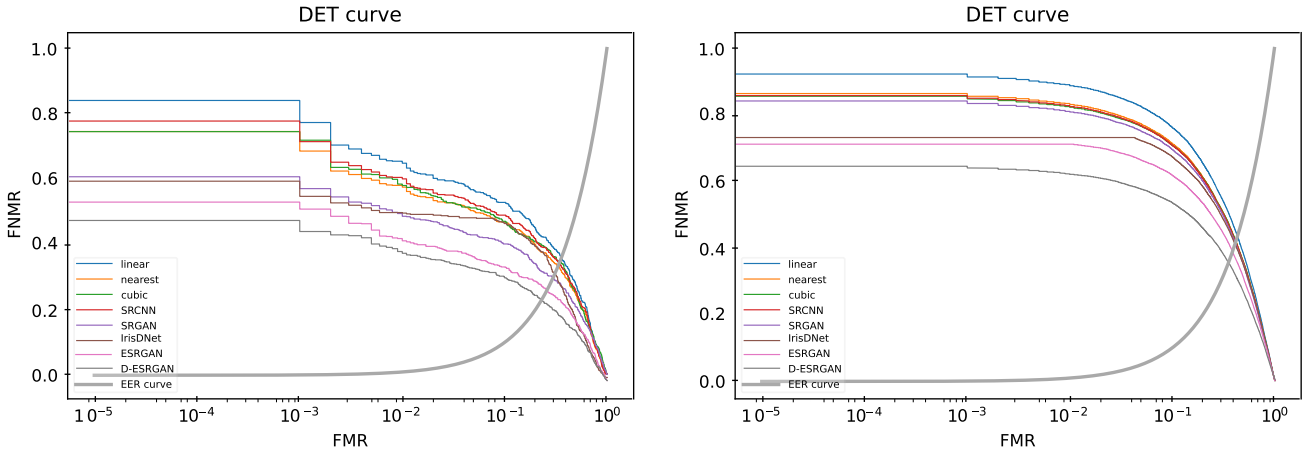


Figure 5. A qualitative comparison of different super-resolution methods on the image quality.



(a) CASIA-Iris-distance (b) CASIA-Iris-M1

Figure 6. DET curves of iris recognition experiments using different iris super-resolution methods.

resolution performance; 2) The use of the Triplet loss can further improve the super-resolution image quality and the iris recognition performance. This demonstrates that the Triplet loss can indeed inject the identity information in training the proposed model, which has an effect on the final recognition results; and 3) the pre-training strategy again improves the super-resolution quality and recognition performance of the proposed model. This is mainly because that pre-training learns a better initialization model parameters, which facilitates the model optimization and prevents overfitting to a certain dataset with limited data.

4.3. Conclusion

This paper proposes a new iris image super-resolution method based on the dual-encoder super-resolution generative adversarial network, which is named as D-ESRGAN. Different from the previous methods, the proposed model integrates both residual CNN encoder and an emerging vision transformer encoder in the generator, which is used to extract local features and capture global associative information, respectively. We also propose a three-stage strategy to better train the model's generator and discriminator from the perspectives of feature reconstruction, image perception, and iris recognition, respectively. Experimental results on the public CASIA-Iris-distance and CASIA-Iris-M1 datasets show that D-ESRGAN archives better performance than state-of-the-art baselines in terms of both super-resolution image quality metrics and iris recognition metric.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No. 62106015, U1836217, 62006225, 62176025), the Beijing University of Civil Engineering and Architecture Research Capacity Promotion Program for Young Scholars (Grant No. X21079) and the Pyramid Talent Training Project of Beijing University of Civil Engineering and Architecture (Grant No. JDYC20220819).

References

[1] K. W. Bowyer and P. J. Flynn. The nd-iris-0405 iris image dataset. *arXiv preprint arXiv:1606.04853*, 2016.

[2] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, 2015.

[3] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision*, pages 391–407, 2016.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Y. Fan, J. Yu, D. Liu, and T. S. Huang. Scale-wise convolution for image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10770–10777, 2020.

[6] X. Gao, K. Zhang, D. Tao, and X. Li. Image super-resolution with sparse neighbor embedding. *IEEE Transactions on Image Processing*, 21(7):3194–3205, 2012.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Proceedings of the Advances in Neural Information Processing Systems*, 27, 2014.

[8] Y. Guo, Q. Wang, H. Huang, X. Zheng, and Z. He. Adversarial iris super resolution. In *Proceedings of the International Conference on Biometrics*, pages 1–8, 2019.

[9] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.

[10] S. Y. Kim, J. Oh, and M. Kim. Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3116–3125, 2019.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] P. Z. Kunszt, A. S. Szalay, and A. R. Thakar. The hierarchical triangular mesh. In *Mining the sky*, pages 631–637. Springer, 2001.

[13] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–632, 2017.

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.

[15] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017.

[16] N. Liu, H. Li, M. Zhang, J. Liu, Z. Sun, and T. Tan. Accurate iris segmentation in non-cooperative environments using fully convolutional networks. In *Proceedings of the International Conference on Biometrics*, pages 1–8. IEEE, 2016.

[17] O. Rukundo and H. Cao. Nearest neighbor value interpolation. *arXiv preprint arXiv:1211.1768*, 2012.

[18] M. S. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4491–4500, 2017.

[19] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[20] B. I. Test. Casia.v4 database. http://www.idealtest.org/dbDetailForUser.do?id=4, Last Accessed (Aug 2022).

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems*, 30, 2017.

[22] C. Wang, Y. Wang, B. Xu, Y. He, Z. Dong, and Z. Sun. A lightweight multi-label segmentation network for mobile iris biometrics. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1006–1010. IEEE, 2020.

[23] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 0–0, 2018.

[24] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers. A fully progressive approach to single-image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 864–873, 2018.

[25] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5790–5799, 2020.

[26] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020.

[27] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19(11):2861–2873, 2010.

[28] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[29] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3929–3938, 2017.

[30] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, pages 286–301, 2018.

[31] Z. Zhao and A. Kumar. Towards more accurate iris recognition using deeply learned spatially corresponding features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3809–3818, 2017.